

Detection of Vandalism in Wikipedia using Metadata Features – Implementation in Simple English and Albanian sections

Arsim Susuri^{*}, Mentor Hamiti², Agni Dika²

¹PhD Student, Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Macedonia

²Professor, Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Macedonia

ARTICLE INFO

Article history:

Received: 22 February, 2017

Accepted: 20 March, 2017

Online: 27 March, 2017

Keywords:

Machine learning

Wikipedia

Vandalism

Metadata features

ABSTRACT

In this paper, we evaluate a list of classifiers in order to use them in the detection of vandalism by focusing on metadata features. Our work is focused on two low resource data sets (Simple English and Albanian) from Wikipedia. The aim of this research is to prove that this form of vandalism detection applied in one data set (language) can be extended into another data set (language). Article views data sets in Wikipedia have been used rarely for the purpose of detecting vandalism. We will show the benefits of using article views data set with features from the article revisions data set with the aim of improving the detection of vandalism. The key advantage of using metadata features is that these metadata features are language independent and simple to extract because they require minimal processing. This paper shows that application of vandalism models across low resource languages is possible, and vandalism can be detected through view patterns of articles.

1. Introduction

Vandalism is a great challenge for Wikipedia, with humans being the main cause, through various illegitimate acts leaving traces in computer systems. Our hypothesis is that vandalism can be characterized through models of article views of vandalized articles in Wikipedia and that vandalism behavior is similar across different languages. In the past, a similar research was done in [1] and [2].

This paper is an extension of work originally presented in [3], by addressing the issue of using metadata features in predicting vandalism in Wikipedia's articles across languages.

According to our hypothesis, a model developed in one language can be applied to other languages. If successful, this would drop the costs of training the classifiers separately for each language.

Applying this model of vandalism detection across different languages shows similar results. In this paper, we will explore the possibility of applying the detection of vandalism across languages

through article views daily and through article editing data set. We combine these data sets in order to analyze any gains in terms of language independency of certain features.

For this purpose, we compare performances of standard classifiers for identifying vandalism in two Wikipedia data sets (Simple English and Albanian). On top of this, we compare the performances of classifiers in one language and the other one and in the combined data set.

2. Approaches

Since 2008 Wikipedia vandalism detection based on machine learning approaches has become a field of increasing research interest. In [4] authors contributed the first machine learning vandalism detection approach using textual features as well as basic metadata features with a logistic regression classifier.

In [5] authors used a Naive Bayes classifier on a bag of words edit representation and were the first to use compression models to detect vandalism in Wikipedia. In [6] authors used Dynamic Markov Compression to detect vandalism edits in Wikipedia.

In [7] author extended the approach in [4] by adding some additional textual features and multiple wordlist-based features. In [8] authors were among the first to present a vandalism

^{*}Corresponding Author: Arsim Susuri, Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Macedonia
Email: arsimsusuri@gmail.com

detection approach based only on spatial and temporal metadata, without the need to inspect article or revision texts.

In [9] authors, in a similar fashion, built a vandalism detection system on top of their WikiTrust reputation system [10]. In [11] authors combined natural language, spatial, temporal and reputation features used in their aforementioned works [7, 8, 9]. Besides in [11], in [12] authors were the first to introduce ex post facto data as features, for whose calculation also future revisions have to be considered. Previously, the process of the detection of vandalism required building a separate learning system for each data set, where by research is focused on each language individually, as in the case in [12]. In the learning process of transfer learning, the domains of the source task and target task are different, as explored in [13]. This work focuses specifically on learning vandalism from one article and applying the models to another article. Supporting the current trend of creating cross language vandalism classifiers, in [2] authors evaluated multiple classifiers based upon a set of language independent features that were compiled from the hourly article view counts and Wikipedia's complete edit history.

We use two data sets from Wikipedia: full history of edits in Wikipedia's Simple English and Albanian data sets¹ and article views on a daily basis². For future reference, we designate data sets as "simple" for Simple English and "sq" for Albanian. The processing of these data sets is explained in the following.

2.1. Wikipedia Data sets

We use the Wikipedia History Dumps of edits dated 29.10.2015, for Simple English and Albanian. In Table 1 are summarized the number of articles and of the revisions, along with usernames. The contents of the articles, used throughout the paper, are encyclopedic and do not include re-direct articles, discussions between users, and other help-related articles.

Table 1: Statistical data of editing history – January–April 2015

Data set	Articles	Revisions	Users
Simple English	413.249	5.565.876	575.755
Albanian	172.150	1.847.827	89.843

The raw data set of article views includes MediaWiki projects, including Wikipedia. In Table 2 are shown statistical data from the raw data set, with article views as filtered data. The filtering process of raw data set is based upon the analysis period of January – April 2015. Although this period of time is relatively short, we demonstrate the viability and success in the detection of vandalism based on models created from within the data set of article views.

Table 2: Statistical data of article views – January–April 2015

Data set	Article Views
Simple English	53.866.869
Albanian	16.698.447

2.2. Revisions with vandalized content

From the main data set, each revision is transformed to a set of features, as shown in Table 3. Feature selection is based on simplicity and language independence and is similar as in [3]. For

each revision, we analyze comments made about it, looking for keywords that might suggest vandalism repair.

If this type of comment is detected, we designate the previous revision as vandalism. The process of labeling revisions is incomplete and noisy. In the past, active learning has been used to solve this issue as in [14].

However, in terms of quality, automatic approach has its limitations, thus requiring the assistance of humans in specific cases of vandalism, as explained in [15]. Based on the reported period of analysis (January – April), we find that approximately 2% of revisions contain vandalism.

This is in consistency with values provided in [16], but less than values reported in [17] and in [18], which report 4-7% of revisions with vandalized content.

2.3. Article views

The raw data set is structured through article views on the hour level. We apply transformation and filtering of articles viewed on the data set, containing previous revisions. The resulting features are shown in Table 4.

In [17] authors showed that these article views are important in order to see the impact of vandalism in Wikipedia. The behavior of the vandals can be analyzed through models, as a result of vandals controlling their work and as a result of increase in curiosity from other users. In [17] authors obtained article views from Wikipedia server logs. This method offered very precise data, in terms of time, but creates a lot of data to be processed.

Many researchers use this data set. A relevant study done in [19] uses this data set in order to compare accesses in medical related information such as allergies. Access models on this data set take into account the impact of seasonal diseases. On the other hand, online users have much more access to Wikipedia than other online medical encyclopedias.

Wikipedia is a well-known on-line source of medical information. Although vandalism has not been included within this study, access models based on seasons do indicate potential vandalism targets.

To determine whether these article views appear at the moment when they were vandalized, we apply search in the edit history data set and label all article views of observed revisions as legitimate or vandalizing. We do not take into account revisions made before January 2015, or articles without any revisions made during the period January – April 2015. This way, we obtain a labeled data set in terms of revisions being vandalized or not.

The final size of data is identical to the size of the combined data set, as explained below. This labeled data set enables us to determine whether or not article view models can predict vandalism. From this final combined data set, we split the time stamp attribute in the hour attribute. With this, we enable machine learning algorithms to learn models of daily access.

2.4. The combined data set

The combined data set is a result of merging two sets of time series for each language. The data set is built by adding features

¹ <https://dumps.wikimedia.org/backup-index.html>
www.astesj.com

² <https://dumps.wikimedia.org/other/pagecounts-raw/>

from the revisions data set labeled in the article views data set by repeating features of revisions.

This way, for each article view, we have information about whether or not the vandalized revision is viewed and, which are the revision features. The process of merging is shown in Figure 1.

We use the split of the “hour” attribute from the time stamp of the article views data set. Based on this split, we obtain eight features in the combined data set: hour, number of requests, transferred information, anonymous edit, minimal revision, size of comment, size of article and vandalism (class label).

These features are language independent and catch metadata of the revisions used more frequently, along with models of access. In order to apply classification algorithms, we split the combined data set to the training data set (January – March) and into the testing data set (April). Statistical data of these data sets are shown in Table 6.

Table 3: Description of features in the edit history data set

Attribute	Description
Title of article	Unique identifier of the Wikipedia article.
Time stamp	Time when the revision was done.
Anonymous edit	Only if an IP address is given. Value 0 is associated with a registered user. Value 1 is associated with an anonymous user.
Minimal revision	Minimal revision has the value 1. Normal revision has the value 0. The editor can emphasize that he/she made minimal revisions in the article (re-formatting, grammar, etc.).
Size of comment	Size of comment (in bytes).
Size of article	Size of the revised article (in bytes).
Vandalism	Revision is classified as vandalism based on the analysis of comments (of the current or upcoming revisions). Value 0 is associated with legitimate revision. Value 1 is associated with vandalizing revision.

Table 4: Description of features in the article views data set

Attribute	Description
Name of the project	Name of the MediaWiki project. In our case, Wikipedia’s Simple English (“simple”) and Albanian (“sq”).
Time stamp	Time stamp of the revision.
Title of article	Title of the Wikipedia article.
Number of requests	Number of requests at a certain hour.
Transferred information	Transfer of data (bytes) from various requests.

2.5. Performance measures

For measuring the efficiency of the classifiers, we will use Area Under Precision-Recall (AUC-PR) and Receiver Operating Characteristic (ROC), as described in [20].

AUC-PR determines the probability that the classifier correctly identifies a random positive sample (vandalism) as positive.

AUC-ROC determines the probability that a classifier correctly identifies a random sample (positive or negative). Both have values ranging from 0 to 1, where value 1 means 100% correctness in labeling all samples taken into consideration. These evaluations are implemented with a confusion matrix, based in [21], as shown in Table 5:

Table 5: Example of confusion matrix

Actual class	Classifier prediction	
	Positive	Negative
Positive	True positive (T_P)	False positive (F_P)
Negative	False negative (F_N)	False positive (F_P)

3. Detection of vandalism across languages

We use the Weka tool, which offers well known machine learning algorithms.

The following supervised machine learning algorithms have been used for this type of vandalism detection:

1. Random Forest (RF) – a supervised classification algorithm [22], which builds the model from many trained decision-making trees from the training data set sample. The default Gini’s impurity criterion is used in order to ascertain the best split on the data feature.
2. Gradient Tree Boosting (GTB) – a supervised algorithm (ensemble tree) based on boosting in order to create a better classifier by optimizing the loss function. Because of the fact that for classification purposes we use two classes, we use the binomial deviation as in [23].
3. Nearest Neighbour (NN) – non parametric classification algorithm. Used in KDTree structures of Bentley [24] because of efficiency in determining separated points and in order to avoid brute force search of the Naïve Nearest Neighbour algorithm.
4. Stochastic Gradient Descent (SGD) – a stochastic approximation to the gradient descent optimization method for minimizing an objective function that is written as a sum of differentiable functions. It is one of the ways of creating linear classifiers. Easily scalable although requires adjusting many parameters. As a result of using too many parameters, the selection of the loss function impacts the performance of the classifiers. As a loss function, we use the logistic regression.

The experiment was conducted in such a way that different arrangements for the above listed classifiers are tested, although with slight differences in results. The reason for the slight differences lies in the fact that all classifiers have converged for a very large number of observations.

If we analyze the data in Table 6, we can conclude that they are not balanced which, in turn, causes problems in the performance of the classifiers used for the experiments. We solve this problem by under sampling the legitimate observations until they match the number of vandalism observations. We extend this application to other data sets.

As a result, we build a set of a balanced subset of the training and testing data. For the detection of vandalism across language, we first train the classification models for the two languages in our

data sets: Simple English and Albanian. Afterwards, these models are evaluated on the testing set for the same language, then on the testing set of the other language.

The similarities between language domains are captured by using metadata, which are language independent. The model of applying the detection of vandalism across languages enables us to generalize the editing and viewing features in Wikipedia.

Application of these models across languages has been successful in the research area of text categorization across languages [25]. In cases of applying these models in text, cultural knowledge of the relevant target language is needed as additional information for the classifiers.

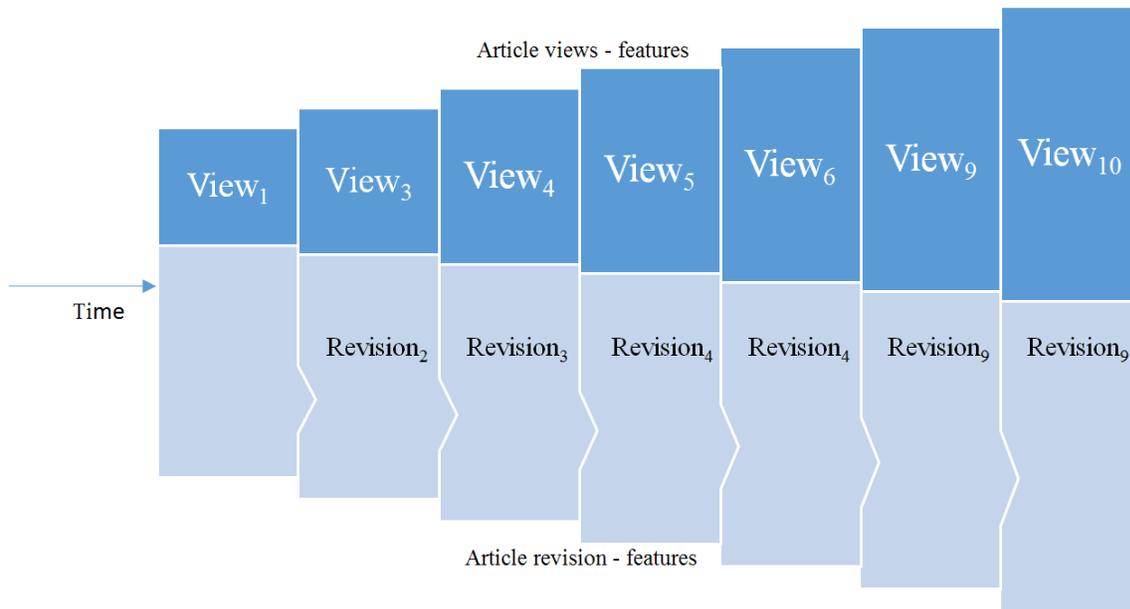


Table 6: Data sets with corresponding vandalism

Data set	Number of articles	Article views	Training set	Testing set
Simple English	413.249	53.866.869	27.543.172	9.611.483
Vandalism	(2.23%) 9.215	-	(2.14%) 589.423	(2.02%) 194.152
Albanian	172.150	16.698.447	9.943.521	5.140.374
Vandalism	(0.12%) 206	-	(0.16%) 15.909	(0.11%) 5654

The advantage of applying these vandalism detection models, built across languages, is that one model can be used for multiple languages, saving resources in developing models for each language. This is especially convenient in Wikipedia, since it contains hundreds of language sections.

This works allows the potential generalization of the concentration of vandalism research in Simple English to other low resource languages, without additional inputs.

4. Experiments and Results

The classification results are shown in Figures 2 to 7. We can see in these Figures the differences in AUC-PR and AUC-ROC values within the same language, and between the two languages used in the experiments. In the case of the designation "simple-sq," the model of the classification is trained in the English language data set and then applied in the testing data set of the Albanian language.

As far as applying classification models in the single language data sets (simple-simple, sq-sq), methods based on trees have better performances, with regards to AUC-PR and AUC-ROC

values. In our case, for the revision data set, higher values have been obtained for GTB and RF, and in the case of the views data set, RF has higher values. However, in terms of time-related costs, these classifiers are the most expensive, as shown in Table 7.

Methods based on trees have higher classification results in the revision data sets.

Applying models across languages obtained lower values, proportionally, although with similar stability in comparison to the application in one language. GTB and RF classifiers have higher values in comparison to other tested classifiers.

SGD classifier has shown better results in single language data sets, and in the case when the training is based in the revision data set of the English language. Based on these results, we can conclude that the English language offers more patterns for detecting vandalism.

If we combine this fact with the fact that SGD is the fastest algorithm for training purposes (Table 7), the benefit, in terms of time costs, is much higher.

In general, the combination of the data sets does not increase the performance of the classifiers (Figures 4 and 7). There is an evident trend of an increase (although slight increase) in the performance of the classifiers of the combined data sets in comparison with the individual data sets.

Based on this, we can conclude that classifiers learn from the best models in each data set (language) but the improvements are not persistent.

Table 7: Execution time (in seconds) for the tested classifiers

		Classifier			
		RF	GTB	NN	SGD
Training set	simple	60	150	12	6
	sq	4	12	2	1
Testing set	simple-simple	11	4	0.5	35
	sq-sq	0.5	0.5	2	0.5
	sq-simple	6	4	65	2
	simple-sq	0.5	0.5	3	0.5

The classification values in the particular language data sets of the revisions (Figures 2 and 5) are at the same level or slightly higher (up to 5%) than the actual systems. Our classification results in particular language data sets have higher AUC-PR values than AUC-ROC (Figures 2, 3, 5 and 6).

In general, the RF classifier is more appropriate for detecting vandalism across languages in terms of cost-related requirements (faster training and testing times).

If we compare AUC-PR values with AUC-ROC values of the RF classifier and GTB classifier, they are on the same level except for the training time (GTB classifier requires much more training time than RF classifier).

Another advantage of the RF classifier is the ability of scalability, which enables parallelism for vandalism detection models on full Wikipedia’s sets.

The advantage of the data sets presented here is in the extraction of language independent features. These features, along with basic classifying algorithms show better performances in comparison to previous studies. The combination of editing and reading patterns shows improvement in the performance of the classifiers and enables these classifiers to use the best features from two data sets in order to predict vandalism. The RF classifier results are comparable to the results obtained in [2].

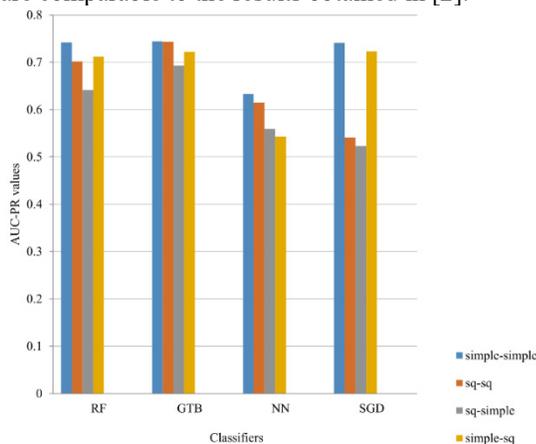


Figure 2: AUC-PR values for the article revisions data set

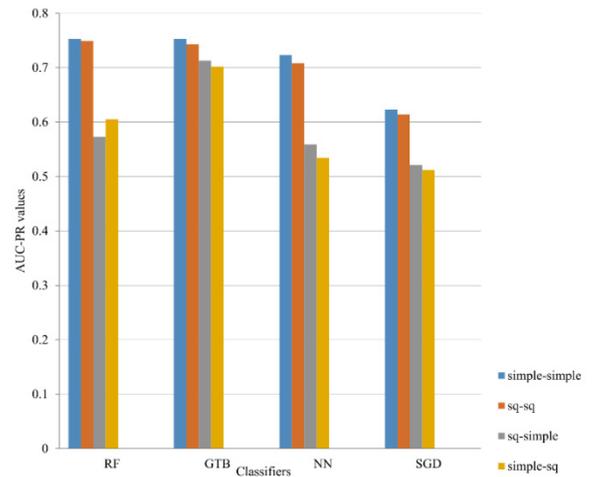


Figure 3: AUC-PR values for the article views data set

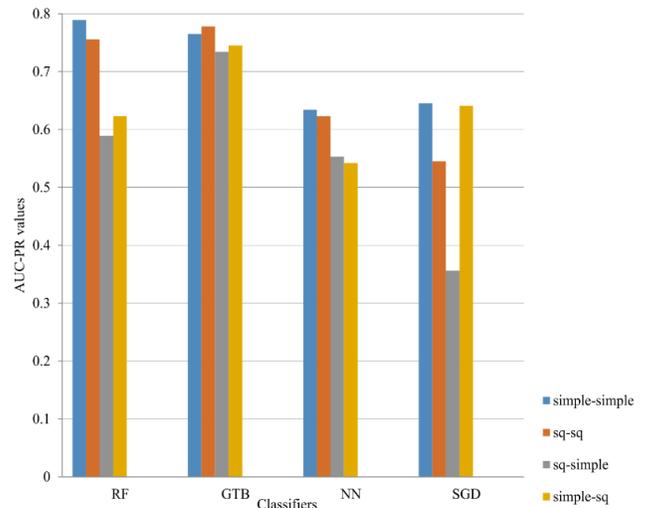


Figure 4: AUC-PR values for the combined data set

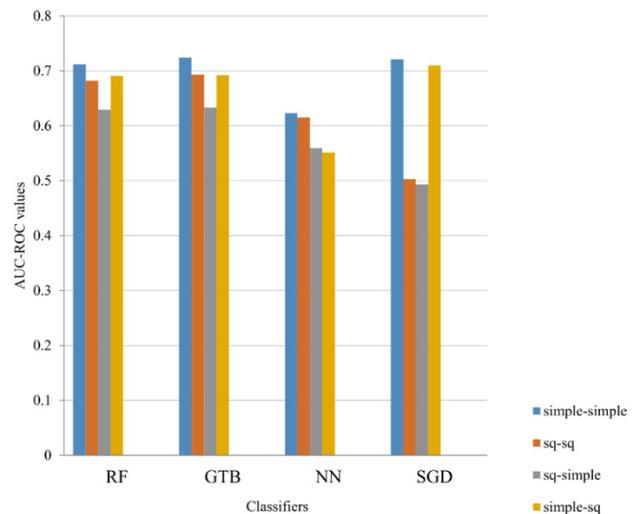


Figure 5: AUC-ROC values for the article revisions data set

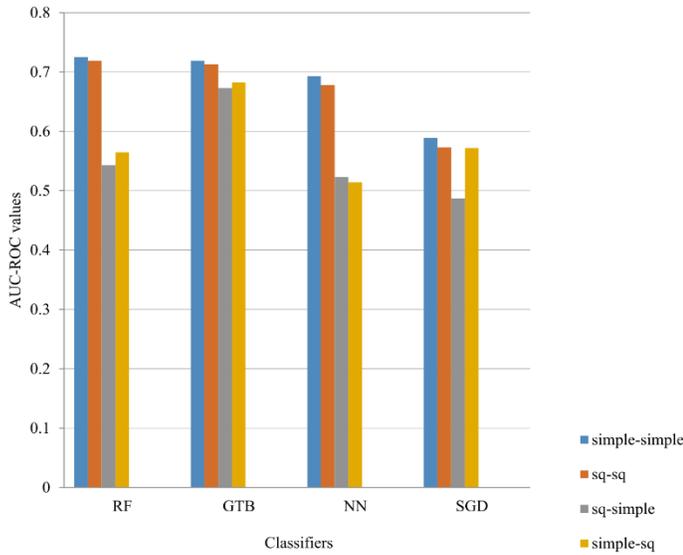


Figure 6: AUC-ROC values for the article views data set

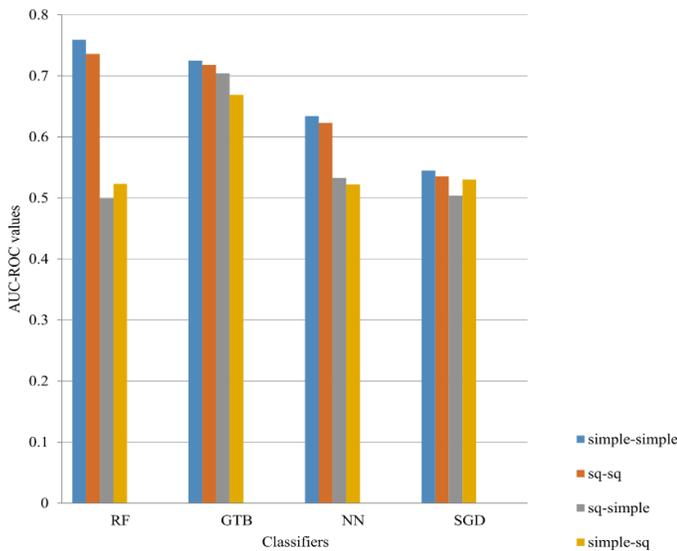


Figure 7: AUC-ROC values for the combined data set

Table 8 shows the impact of the features from the article views data set for the performance of the RF classifier. In this table, we have summarized the comparisons of the two data sets in terms of rankings of the features, based on the RF classifier. It comparatively describes the impact that each feature has on the overall performance of the RF classifier.

Although the improvements are not high, the addition of the features from the article views data set does not have a negative impact on the overall performance. The article views data set alone is not sufficient for vandalism detection and requires labeling from the revisions' data set.

However, the article views data set is a simple data set with few features that show some changes in access patterns when vandalism has occurred.

Table 8: Feature rankings of the combined data set based on the RF classifier

Features	Combined Data sets	
	simple	sq
Size of comment	0.358	0.243
Transferred bytes	0.261	0.238
Number of requests	0.196	0.354
Minimal revision	0.086	0.089
Anonymous edit	0.042	0.036
Size of article	0.041	0.024
Hour	0.017	0.017

5. Conclusions and Future Work

In this paper, we have presented data sets for the detection of vandalism and have demonstrated the application of four machine learning algorithms for the detection of vandalism within different languages and across languages. We have created three data sets from the data set of article views; full history of article edits and their combination. We have analyzed two Wikipedia editions: Simple English and Albanian.

During the experimentations, we have found out that the GTB classifier showed better performances in predicting vandalism, although in terms of time, it has higher costs.

The RF classifier has similar performances (0.2% - 0.5% difference) in comparison to the GTB classifier but with very low training costs (Table 7).

These results show that viewing and editing features of vandals are similar across languages. As a result of this fact, vandalism models of one language can be trained in one language and applied into another language. We have shown that application of the vandalism model across languages is feasible, and that view patterns can be used to detect and predict vandalism.

For future research, the inclusion of popular articles and the changes in traffic, caused by the vandalism, would be the right step to better understanding of the correlation of different data sets with regards to the impact they have on improving the vandalism detection rates.

References

- [1] G. West. "Damage Detection and Mitigation in Open Collaboration Applications," Ph.D. Thesis, University of Pennsylvania, 2013.
- [2] K. N. Tran, P. Christen, "Cross-language prediction of vandalism on wikipedia using article views and revisions", Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2013.
- [3] A. Susuri, M. Hamiti and A. Dika, "Machine Learning Based Detection of Vandalism in Wikipedia across Languages" in proceedings of the 5th Mediterranean Conference on Embedded Computing (MECO), Bar, Montenegro, 2016.
- [4] M. Potthast, B. Stein, and R. Gerling, "Automatic vandalism detection in wikipedia" in advances in information retrieval, 663-668. Springer Berlin Heidelberg, 2008.
- [5] K. Smets, B. Goethals, and B. Verdonk, "Automatic vandalism detection in wikipedia: Towards a machine learning approach" in WikiAI '08: Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence, 2008.
- [6] K. Y. Itakura and C. L. a. Clarke, "Using dynamic markov compression to detect vandalism in the Wikipedia" Proceedings of the 32nd international

ACM SIGIR conference on Research and development in information retrieval - SIGIR '09, 822, 2009.

- [7] S. M. Mola-Velasco, "Wikipedia vandalism detection through machine learning: Feature review and new proposals - lab report for pan at clef 2010" in CLEF (Notebook Papers/LABs/Workshops), 2010.
- [8] A. G. West, S. Kannan, and I. Lee, "Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata" in Proceedings of the Third European Workshop on System Security, EUROSEC '10, 22-28, New York, NY, USA, 2010.
- [9] B. T. Adler, L. De Alfaro, and I. Pye, "Detecting wikipedia vandalism using wikitrust" Notebook papers of CLEF, 2010.
- [10] B. T. Adler and L. De Alfaro, "A content-driven reputation system for the Wikipedia" Proceedings of the 16th international conference on World Wide Web WWW 07, 7(Generic):261, 2007.
- [11] B. T. Adler, L. De Alfaro, S. M. Mola-Velasco, Paolo Rosso, and Andrew G. West, "Wikipedia vandalism detection: Combining natural language, metadata, and reputation features" in proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II, 277-288, Berlin, Heidelberg, 2011.
- [12] A. G. West and I. Lee, "Multilingual vandalism detection using language-independent & ex post facto evidence - notebook for pan at clef 2011" in CLEF (Notebook Papers/Labs/Workshop), 2011.
- [13] S. C. Chin and W. N. Street, "Divide and Transfer: an Exploration of Segmented Transfer to Detect Wikipedia Vandalism", Journal of Machine Learning Research (JMLR): Workshop and Conference Proceedings, 27, 133-144, 2012.
- [14] S. C. Chin, W. N. Street, P. Srinivasan, and D. Eichmann, "Detecting Wikipedia Vandalism with Active Learning and Statistical Language Models", in Proceedings of the 4th Workshop on Information Credibility (WICOW), 2010.
- [15] Q. Wu, D. Irani, C. Pu, and L. Ramaswamy, "Elusive Vandalism Detection in Wikipedia: A Text Stability-based Approach" in Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 2010.
- [16] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi, "He says, she says: conflict and coordination in Wikipedia" in Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), San Jose, CA, pages 453 - 462, 2007.
- [17] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl, "Creating, destroying, and restoring value in Wikipedia" in Proceedings of the International ACM Conference on Supporting GroupWork (GROUP), Sanibel Island, FL, pages 259 - 268, 2007.
- [18] M. Potthast, "Crowdsourcing a Wikipedia Vandalism Corpus" in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2010.
- [19] M. R. Laurent and T. J. Vickers, "Seeking Health Information Online: Does Wikipedia Matter?" Journal of the American Medical Informatics Association (JAMIA), 16 (2009), pages 471 - 479, 2009.
- [20] J. Davis, and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves" in: ICML '06: Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania: ACM, pp. 233-240, 2006.
- [21] R. Kohavi, F. Provost, "Glossary of terms, Machine Learning" Vol. 30, No. 2/3, pages 271 -274, 1998.
- [22] L. Breiman, "Random forests" Machine Learning, 45(1):5-32, October 2001.
- [23] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine" Annals of Statistics, 1189-1232, 2001
- [24] J. L. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching" Communications of the ACM, 18, pages 509 - 517, 1975.
- [25] L. Rigutini, M. Maggini, and B. Liu, "An EM Based Training Algorithm for Cross-Language Text Categorization" in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2005.